Multi-view learning for multivariate performance measures optimization

Jim Jing-Yan Wang

***Abstract*—In this paper, we propose the problem of optimizing multivariate performance measures from multi-view data, and an effective method to solve it. This problem has two features: the data points are presented by multiple views, and the target of learning is to optimize complex multivariate performance measures. We propose to learn a linear discriminant functions for each view, and combine them to construct a overall multivariate mapping function for mult-view data. To learn the parameters of the linear discriminant functions of different views to optimize multivariate performance measures, we formulate a optimization problem. In this problem, we propose to minimize the complexity of the linear discriminant functions of each view, encourage the consistences of the responses of different views over the same data points, and minimize the upper boundary of a given multivariate performance measure. To optimize this problem, we employ the cutting-plane method in an iterative algorithm. In each iteration, we update a set of constrains, and optimize the mapping function parameter of each view one by one.**

***Index Terms*—Multivariate performance measures, Multi-view learning, Cutting-plane algorithm**

I. Introduction

DIFFERENT multivariate performance measures are used to evaluate different machine learning applications [1]. For example, in problems of text classification, F1-score and precision/recall breakeven point (PRBEP) are used to compare true class labels against predicted class labels of a given test text set. In image retrieval problems, the precision/recall at the top k returned images are used to evaluate the performance of a retrieval system. Recently, the problem of multivariate performance measures optimization are proposed to learn directly to losses based on pre-defined multivariate performance measures. Many algorithms have been proposed to solve this problem, and some examples are listed as follows.

- Joachims [1] proposed a support vector machine (SVM)-based for multivariate nonlinear performance measures optimization problems. The proposed algorithm can train a multivariate SVM in polynomial time for large classes of potentially non-linear performance measures. Moreover, the traditional SAM can be arised as a special case of this method.
- Li et al. [2] proposed a two-step approach to optimize multivariate performance measures, by first training a nonlinear classifiers with existing learning methods, and then adapting it to optimize specific performance measures. In the seconde step, the classifier adaptation problem can be solved as a quadratic program problem, in a similar way to linear SVM.
- Mao and Tsang [3] proposed a novel feature selection method to optimize multivariate performance measures, by formulating the problem or high-dimensional data, and employing a two-layer cutting plane algorithm to solve it. Moreover, this method is also used in multiple-instance learning problems.

Up to now, all these multivariate methods are limited to learn from data with single view. However, in many real-world machine learning problems, the data can be presented by multiple views. For example, in computer vision problems, we can extract different types of features, such as color features, texture features, and shape features, and each feature can be treated as a view. In scientific article classification problems, we can also learn from the views of article abstract, content, and references. Different views of data may be complemental to each other in a learning problem and using multiple views has been a popular strategy in machine learning community. To learn from multiple views of data, different multi-view learning methods have been proposed [4], [5], [6], [7]. However, none of them are designed to optimized a specific multivariate performance measure.

To overcome this problem, in this paper, we propose the problem of learning from multiple view data to optimize the multivariate performance measures. Given a tuple of data points, each of them are presented with multiple views. The problem is to learn a multivariate mapping function to map them to a tuple of class labels, so that the multivariate performance measures can be optimized. To solve this problem, we proposed a novel method for multi-view learning to optimize multivariate performance measures. The contribution of this paper are of two folds:

1) We propose the problem of multi-view learning for multivariate performance measures. Although there are plenty of multivariate performance measures optimization methods, they are limited to single view data. There are also lots of multi-view learning methods, however, none of them are proposed to optimize multivariate performance measures.
2) We also propose a novel method to solve this problem. We proposed to learn a linear discriminant functions for each view, and combine them to construct a overall multivariate mapping function to predict the class label tuple of a tuple of data points. To learn the linear discriminant functions parameters of different views, we formulate a constrained minimization problem. In this problem, we proposed to minimize the complexity of each linear discriminant functions parameter by minimizing its squared ℓ_2 norm, encourage the consistence among different views by minimizing the squared ℓ_2 norm distance

among different linear discriminant functions response of differen views, and also minimize the losses based on a specific multivariate performance measures. The minimization problem is optimized by a cutting-plane method in an alterative algorithm [8]. We maintain a active set of constrains, and update it in each iteration by add a most violated class label tuple. We also optimize the multivariate mapping function parameters one by one, and the optimization of a multivariate mapping function parameter can be solve as a quadratic program problem.

The rest parts of this paper are organized as follows: in section II, we introduce the proposed method, and in section III, we conclude this paper.

II. Proposed method

A. Problem formulation

We assume we have a training data set of n data points, and each data point has m views. The training set is denoted as $\{(\mathbf{x}_i^j|_{j=1}^m, y_i)\}|_{i=1}^n$, where $\mathbf{x}_i^j \in \mathbb{R}^{d_j}$ is the d_j-dimensional feature vector of the j-th view of the i-th data point, and $y_i \in \{+1, -1\}$ is the binary class label of the i-th data point. We propose to learn a multivariate mapping function $\overline{h}$ to map a tuple of n data points of m views, $\overline{\mathbf{x}} = (\mathbf{x}_1^j|_{j=1}^m, \cdots, \mathbf{x}_n^j|_{j=1}^m)$, to tuple of n class labels, $\overline{y} = (y_1, \cdots, y_n)$. To implement this multivariate mapping function, we use a linear discriminant function $f_j(\overline{\mathbf{x}}^j, \overline{y}')$ to predict the response of a view the the data tuple, $\overline{\mathbf{x}}^j = (\mathbf{x}_1^j, \cdots, \mathbf{x}_n^j)$, against a candidate class label tuple $\overline{y}' = (y_1', \cdots, y_n')$,

$$f_j(\overline{\mathbf{x}}^j, \overline{y}') = \sum_{i=1}^n \mathbf{w}_j^\top \Psi(\overline{\mathbf{x}}^j, \overline{y}') \tag{1}$$

where $\mathbf{w}_j \in \mathbb{R}^{d_j}$ is a parameter vector for the linear discriminant function of the j-th view, , and $\Psi(\overline{\mathbf{x}}^j, \overline{y})$ is a function which can returns a vector to describe the match between $\overline{\mathbf{x}}^j$ and $\overline{y}'$, defined as follows,

$$\Psi(\overline{\mathbf{x}}^j, \overline{y}') = \sum_{i=1}^n y_i' \mathbf{x}_i^j. \tag{2}$$

Based on the linear discriminant functions of different views, we construct the multivariate mapping function as follows,

$$\overline{h}(\overline{\mathbf{x}}) = \arg\max_{\overline{y}' \in \mathcal{Y}} \left\{ \sum_{j=1}^m f_j(\overline{\mathbf{x}}^j, \overline{y}') = \sum_{j=1}^m \mathbf{w}_j^\top \Psi(\overline{\mathbf{x}}^j, \overline{y}') \right\} \tag{3}$$

where $\mathcal{Y} = \{+1, -1\}^n$ is a set of all admissible label vectors.

We propose to optimize a complex multivariate performance measure, Δ, by learning the parameter vectors $\mathbf{w}_j|_{j=1}^m$ for the m views. To this end, we consider the following three problems,

1) **Reducing the complexity of each linear discriminative function**: To prevent over-fitting, we propose to reduce the complexity of the linear discriminative function of each view by minimizing the squared ℓ_2 norm of its parameter,

$$\min_{\mathbf{w}_j|_{j=1}^m} \frac{1}{2} \sum_{j=1}^m \|\mathbf{w}_j\|_2^2. \tag{4}$$

2) **Encouraging consistences among different views**: To encourage the consistences of different views, we proposed to minimize the squared ℓ_2 norm distances of responses of any two linear discriminative functions over one single data point,

$$\min_{\mathbf{w}_j|_{j=1}^m} \frac{1}{2} \sum_{j,j':j'<j} \left(\sum_{i=1}^n \|\mathbf{w}_j^\top \mathbf{x}_i^j - \mathbf{w}_{j'}^\top \mathbf{x}_i^{j'}\|_2^2 \right). \tag{5}$$

3) **Optimizing multivariate performance measures**: To optimize a specific multivariate performance measure, we propose to minimize the upper boundary of a loss function Δ based on this multivariate performance measure,

$$\begin{aligned} &\min_{\mathbf{w}_j|_{j=1}^m, \xi} \xi \\ &\quad s.t. \forall \overline{y}' \in \mathcal{Y}/\overline{y}: \\ &\qquad \sum_{j=1}^m \mathbf{w}_j^\top \left[\Psi(\overline{x}^j, \overline{y}) - \Psi(\overline{x}^j, \overline{y}')\right] \geq \Delta(\overline{y}', \overline{y}) - \xi, \end{aligned} \tag{6}$$

where ξ is a slack variable of the upper boundary of the loss function.

The overall optimization function are obtained by combining all the problems above,

$$\begin{aligned} &\min_{\mathbf{w}_j|_{j=1}^m, \xi} \left\{ \frac{1}{2} \sum_{j=1}^m \|\mathbf{w}_j\|_2^2 + C_1 \xi \right. \\ &\qquad \left. + \frac{C_2}{2} \sum_{j,j':j'<j} \left(\sum_{i=1}^n \|\mathbf{w}_j^\top \mathbf{x}_i^j - \mathbf{w}_{j'}^\top \mathbf{x}_i^{j'}\|_2^2 \right) \right\} \\ &s.t. \forall \overline{y}' \in \mathcal{Y}/\overline{y}: \\ &\quad \sum_{j=1}^m \mathbf{w}_j^\top \left[\Psi(\overline{x}^j, \overline{y}) - \Psi(\overline{x}^j, \overline{y}')\right] \geq \Delta(\overline{y}', \overline{y}) - \xi. \end{aligned} \tag{7}$$

In the objective of this problem, the first term is to reduce the complexity of the parameter vector of each view, and second term is a slack variable to represent the upper boundary of the multivariate performance measure, and the third term is to encourage the consistences of the responses of different views. C_1 and C_2 are tradeoff parameters.

B. Problem optimization

To solve this problem, we employ the cutting-plane algorithm. In an iterative algorithm, we update the parameter vectors $\mathbf{w}_j|_{j=1}^m$ and an active set of constrains $\mathcal{W}$ alternately.

1) Updating $\mathbf{w}_j|_{j=1}^m$: When we have a given active set of constrain $\mathcal{W} \subseteq \mathcal{Y}/\overline{y}$, and optimize the parameter vectors, we optimize $\mathbf{w}_j$ one by one, i.e., when $\mathbf{w}_j$ is optimized, $\mathbf{w}_{j'}|_{j'\neq j}$ is fixed. The following optimization is obtained in this case,

$$
\begin{aligned}
&\min_{\mathbf{w}_j,\xi} \left\{ \frac{1}{2}\|\mathbf{w}_j\|_2^2 + C_1\xi \right.\\
&\qquad \left. + \frac{C_2}{2}\sum_{j':j'<j}\left(\sum_{i=1}^n \|\mathbf{w}_j^\top \mathbf{x}_i^j - \mathbf{w}_{j'}^\top \mathbf{x}_i^{j'}\|_2^2\right)\right\}\\
&s.t. \forall \overline{y}' \in \mathcal{W}:\\
&\quad \mathbf{w}_j^\top\left[\Psi(\overline{x}^j,\overline{y}) - \Psi(\overline{x}^j,\overline{y}')\right]\\
&\quad \geq \Delta(\overline{y}',\overline{y}) - \sum_{j':j'\neq j}\mathbf{w}_{j'}^\top\left[\Psi(\overline{x}^{j'},\overline{y}) - \Psi(\overline{x}^{j'},\overline{y}')\right] - \xi.
\end{aligned}
\tag{8}
$$

The Lagrange function of this problem is

$$
\begin{aligned}
&\mathcal{L}(\mathbf{w}_j,\xi,\alpha_{\overline{y}'}|_{\overline{y}':\overline{y}'\in\mathcal{W}})\\
=&\frac{1}{2}\|\mathbf{w}_j\|_2^2 + C_1\xi\\
&+\frac{C_2}{2}\sum_{j':j'<j}\left(\sum_{i=1}^n \|\mathbf{w}_j^\top \mathbf{x}_i^j - \mathbf{w}_{j'}^\top \mathbf{x}_i^{j'}\|_2^2\right)\\
&-\sum_{\overline{y}':\overline{y}'\in\mathcal{W}}\alpha_{\overline{y}'}\left(\mathbf{w}_j^\top\left[\Psi(\overline{x}^j,\overline{y}) - \Psi(\overline{x}^j,\overline{y}')\right] - \Delta(\overline{y}',\overline{y})\right.\\
&\left.+\sum_{j':j'\neq j}\mathbf{w}_{j'}^\top\left[\Psi(\overline{x}^{j'},\overline{y}) - \Psi(\overline{x}^{j'},\overline{y}')\right] + \xi\right)\\
=&\frac{1}{2}\mathbf{w}_j^\top\Omega\mathbf{w}_j + C_1\xi - \mathbf{w}_j^\top\boldsymbol{\beta} + \frac{C_2}{2}\sum_{j':j'<j}\sum_{i=1}^n \mathbf{w}_{j'}^\top\mathbf{x}_i^{j'}\mathbf{x}_i^{j'\top}\mathbf{w}_{j'}\\
&-\sum_{\overline{y}':\overline{y}'\in\mathcal{W}}\alpha_{\overline{y}'}\left(\mathbf{w}_j^\top\boldsymbol{\gamma}_{\overline{y}'} - \delta_{\overline{y}'} + \xi\right)
\end{aligned}
\tag{9}
$$

where $\alpha_{\overline{y}'} \geq 0$ is the Lagrange multiplier for the constrain $\mathbf{w}_j^\top\left[\Psi(\overline{x}^j,\overline{y}) - \Psi(\overline{x}^j,\overline{y}')\right] \geq \Delta(\overline{y}',\overline{y}) - \sum_{j':j'\neq j}\mathbf{w}_{j'}^\top\left[\Psi(\overline{x}^{j'},\overline{y}) - \Psi(\overline{x}^{j'},\overline{y}')\right] - \xi$, and

$$
\begin{aligned}
&\Omega = \left(I + C_2\sum_{j':j'<j}\sum_{i=1}^n \mathbf{x}_i^j\mathbf{x}_i^{j\top}\right),\\
&\boldsymbol{\beta} = C_2\sum_{j':j'<j}\sum_{i=1}^n \mathbf{x}_i^j\mathbf{x}_i^{j'\top}\mathbf{w}_{j'},\\
&\boldsymbol{\gamma}_{\overline{y}'} = \left[\Psi(\overline{x}^j,\overline{y}) - \Psi(\overline{x}^j,\overline{y}')\right]\\
&\boldsymbol{\delta}_{\overline{y}'} = \left(\Delta(\overline{y}',\overline{y}) - \sum_{j':j'\neq j}\mathbf{w}_{j'}^\top\left[\Psi(\overline{x}^{j'},\overline{y}) - \Psi(\overline{x}^{j'},\overline{y}')\right]\right).
\end{aligned}
\tag{10}
$$

This optimization problem can be transferred to its dual form,

$$
\begin{aligned}
&\max_{\alpha_{\overline{y}'}|_{\overline{y}':\overline{y}'\in\mathcal{W}}}\ \min_{\mathbf{w}_j,\xi}\ \mathcal{L}(\mathbf{w}_j,\xi,\alpha_{\overline{y}'}|_{\overline{y}':\overline{y}'\in\mathcal{W}})\\
&s.t.\ \forall\overline{y}',\overline{y}'\in\mathcal{W}:\ \alpha_{\overline{y}'}\geq 0.
\end{aligned}
\tag{11}
$$

To obtain the optimal points of $\mathbf{w}_j$ and ξ, we set the gradients of Lagrangian function with respect to $\mathbf{w}_j$ and ξ to zeros, and we have,

$$
\begin{aligned}
&\frac{\partial\mathcal{L}}{\partial\mathbf{w}_j} = \Omega\mathbf{w}_j - \boldsymbol{\beta} - \sum_{\overline{y}':\overline{y}'\in\mathcal{W}}\alpha_{\overline{y}'}\boldsymbol{\gamma}_{\overline{y}'} = 0\\
&\Rightarrow \mathbf{w}_j = \Omega^{-1}\left(\boldsymbol{\beta} + \sum_{\overline{y}':\overline{y}'\in\mathcal{W}}\alpha_{\overline{y}'}\boldsymbol{\gamma}_{\overline{y}'}\right),\\
&\frac{\partial\mathcal{L}}{\partial\xi} = C_1 - \sum_{\overline{y}':\overline{y}'\in\mathcal{W}}\alpha_{\overline{y}'} = 0\\
&\Rightarrow \sum_{\overline{y}':\overline{y}'\in\mathcal{W}}\alpha_{\overline{y}'} = C_1.
\end{aligned}
\tag{12}
$$

By substituting these results back to (11), we obtain the dual problem as

$$
\begin{aligned}
&\max_{\alpha_{\overline{y}'}|_{\overline{y}':\overline{y}'\in\mathcal{W}}}\left\{-\frac{1}{2}\left(\boldsymbol{\beta} + \sum_{\overline{y}':\overline{y}'\in\mathcal{W}}\alpha_{\overline{y}'}\boldsymbol{\gamma}_{\overline{y}'}\right)^\top\Omega^{-1}\left(\boldsymbol{\beta} + \sum_{\overline{y}':\overline{y}'\in\mathcal{W}}\alpha_{\overline{y}'}\boldsymbol{\gamma}_{\overline{y}'}\right)\right.\\
&\qquad\left. - \sum_{\overline{y}':\overline{y}'\in\mathcal{W}}\alpha_{\overline{y}'}\delta_{\overline{y}'}\right\}\\
&\quad s.t.\ \forall\overline{y}',\overline{y}'\in\mathcal{W}:\ \alpha_{\overline{y}'}\geq 0,\ and\ \sum_{\overline{y}':\overline{y}'\in\mathcal{W}}\alpha_{\overline{y}'} = C_1.
\end{aligned}
\tag{13}
$$

We can solve this problem as a quadratic program problem.

2) Updating $\mathcal{W}$: To update $\mathcal{W}$, we first find the most violated $\overline{y}'$ and then add it the $\mathcal{W}$. $\overline{y}'$ is obtained as

$$
\begin{aligned}
\overline{y}' = \arg\max\nolimits_{\overline{y}''\in\mathcal{Y}/y}&\left\{\Delta(\overline{y}'',\overline{y}) + \sum_{j=1}^m \mathbf{w}_j^\top\Psi(\overline{x}^j,\overline{y}'')\right.\\
&\left. - \sum_{j=1}^m \mathbf{w}_j^\top\Psi(\overline{x}^j,\overline{y})\right\}.
\end{aligned}
\tag{14}
$$

3) Alterative algorithm: The proposed iterative algorithm is given in Algorithm 1.

III. Conclusion

In this paper, we propose the problem of multi-view learning for multivariate performance measures, and a novel algorithm to solve it. Multivariate performance measures optimization can obtain a predictor which directly optimize a desired performance measure. However, the existing methods are limited to single view data, and cannot utilize multi-view data. We propose to learn from multi-view data to optimize the multivariate performance measures, and it has potential in real-world applications. The proposed method can be implemented easily due to its employment of cutting plane method and quadratic program.

Algorithm 1 Iterative multi-view learning algorithm for multivariate performance measures.

Input: Multi-view training data set $\{(\mathbf{x}_i^j|_{j=1}^m, y_i)\}|_{i=1}^n$;
Input: Tradeoff parameters C_1 and C_2;
Input: The maximum iteration number T.
Initialize linear discriminative function parameter $\mathbf{w}_j^0$ for each view;
$\mathcal{W} \leftarrow \emptyset$;
Calculate Ω as in (10);
for $t = 1, \cdots, T$ **do**
  Find the most violated class label tuple $\overline{y}'^t$ as in (14) by fixing $\mathbf{w}_j^{t-1}|_{j=1}^m$;
  Add $\overline{y}'^t$ to the constrain active set $\mathcal{W}$, $\mathcal{W} \leftarrow \mathcal{W} \cup \{\overline{y}'^t\}$;
  for $j = 1, \cdots, m$ **do**
    Update $\mathbf{w}_j^t$ by solving (13) by fixing $\mathbf{w}_{j'}^{t-1}|_{j' \neq j}$ and $\mathcal{W}$;
  end for
end for
Output: $\mathbf{w}_j^T|_{j=1}^m$.

REFERENCES

[1] T. Joachims, "A support vector method for multivariate performance measures," in *Proceedings of the 22nd international conference on Machine learning*. ACM, 2005, pp. 377–384.

[2] N. Li, I. W. Tsang, and Z.-H. Zhou, "Efficient optimization of performance measures by classifier adaptation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 6, pp. 1370–1382, 2013.

[3] Q. Mao and I.-H. Tsang, "A feature selection method for multivariate performance measures," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 9, pp. 2051–2063, 2013.

[4] V. Sindhwani and D. S. Rosenberg, "An rkhs for multi-view learning and manifold co-regularization," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 976–983.

[5] S. Z. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum, "Statistical learning of multi-view face detection," in *Computer VisionłECCV 2002*. Springer, 2002, pp. 67–81.

[6] C. Christoudias, R. Urtasun, and T. Darrell, "Multi-view learning in the presence of view disagreement," *arXiv preprint arXiv:1206.3242*, 2012.

[7] W. Li, L. Duan, I. W.-H. Tsang, and D. Xu, "Co-labeling: A new multi-view learning approach for ambiguous problems." in *ICDM*, 2012, pp. 419–428.

[8] J. E. Kelley, Jr, "The cutting-plane method for solving convex programs," *Journal of the Society for Industrial & Applied Mathematics*, vol. 8, no. 4, pp. 703–712, 1960.